Received 20 December 2022; revised 13 November 2023 and 5 January 2024; accepted 5 February 2024. Date of publication 13 February 2024; date of current version 11 June 2024. The review of this article was arranged by Editor Yudong Zhang.

Digital Object Identifier 10.1109/OJEMB.2024.3365290

Deep Learning-Based Glucose Prediction Models: A Guide for Practitioners and a Curated Dataset for Improved Diabetes Management

SAÚL LANGARICA [1] (Student Member, IEEE), DIEGO DE LA VEGA [2], NAWEL CARIMAN [1], MARTÍN MIRANDA [2], DAVID C. ANDRADE [3], FELIPE NÚÑEZ [1] (Senior Member, IEEE), AND MARIA RODRIGUEZ-FERNANDEZ [2] (Member, IEEE)

[1]Department of Electrical Engineering, Pontificia Universidad Católica de Chile, Santiago 7820436, Chile
[2]Institute for Biological and Medical Engineering, Schools of Engineering, Medicine and Biological Sciences, Pontificia Universidad Católica de Chile, Santiago 7820436, Chile
[3]Centro de Investigación en Fisiología y Medicina de Altura, Facultad de Ciencias de la Salud, Universidad de Antofagasta, Antofagasta 1271155, Chile

CORRESPONDING AUTHOR: MARIA RODRIGUEZ-FERNANDEZ (e-mail: marodriguezf@uc.cl)

This work was supported by the Chilean National Agency for Research and Development (ANID) under Grant ACT210083 and Grant Fondecyt 1230844.

ABSTRACT Accurate short- and mid-term blood glucose predictions are crucial for patients with diabetes struggling to maintain healthy glucose levels, as well as for individuals at risk of developing the disease. Consequently, numerous efforts from the scientific community have focused on developing predictive models for glucose levels. This study harnesses physiological data collected from wearable sensors to construct a series of data-driven models based on deep learning approaches. We systematically compare these models to offer insights for practitioners and researchers venturing into glucose prediction using deep learning techniques. Key questions addressed in this work encompass the comparison of various deep learning architectures for this task, determining the optimal set of input variables for accurate glucose prediction, comparing population-wide, fine-tuned, and personalized models, and assessing the impact of an individual's data volume on model performance. Additionally, as part of our outcomes, we introduce a meticulously curated dataset inclusive of data from both healthy individuals and those with diabetes, recorded in free-living conditions. This dataset aims to foster research in this domain and facilitate equitable comparisons among researchers.

INDEX TERMS Diabetes, Glucose prediction, deep learning, transfer learning.

IMPACT STATEMENT Application of deep learning models to wearable sensor data provides a promising avenue for accurate blood glucose prediction, offering a practical solution for patients managing diabetes and those at risk.

I. INTRODUCTION

In the last decades, diabetes has emerged as a significant global public health concern. The number of individuals living with diabetes increased from 108 million in 1980 to 422 million in 2014, and the global prevalence (age-normalized) nearly doubled from 4.7% to 8.5% among adults during the same period [1]. As of 2019, an estimated 463 million adults were living with diabetes, projected to reach 578 by 2030 and 700 million by 2045 [2].

Individuals with diabetes mellitus face the daily challenge of regulating their blood glucose levels within a healthy range to avoid severe adverse effects associated with both hypoglycemia and hyperglycemia events. While the target range should be personalized to meet each patient's specific

needs [3], the standard target for individuals with Type I Diabetes Mellitus (T1DM) is 70–180 mg/dl. Achieving this range involves weight management, regular physical activity, dietary control, and insulin administration. The latter depends on continuous monitoring of blood glucose levels [4].

Recent technological advances have greatly benefited diabetes treatment. Several mechanisms are now available for supplying exogenous insulin to the body, including insulin syringes, pens, oral medications, and insulin pumps [5]. Furthermore, the advent of Continuous Glucose Monitors (CGMs) has ushered in the era of artificial pancreas (AP) systems, capable of automatically dispensing insulin to enable precise and soft automatic control [6]. In this context, short- and mid-term predictions of blood glucose concentration have become a valuable tool to support individuals with diabetes and those at risk of developing the condition in their efforts to manage and regulate their glucose levels. Notably, the most successful control technique used in AP systems, model-based predictive control (MPC) [7], relies on predictive models that forecast future blood glucose values to calculate the optimal amount of insulin to be supplied.

Several modeling techniques have been assessed for building predictive models of blood glucose, typically with prediction horizons ranging from 30 to 60 minutes into the future [8]. However, this is a challenging task due to the need to account for various disturbances, such as the glycemic impact of exercise (which can lower blood sugar levels) and food intake (which can raise them) [6]. For instance, many commercially available AP systems require user-initiated insulin bolus administration at mealtime [9]. Moreover, understanding how exercise, circadian rhythms, and dietary intake affect blood glucose levels in both healthy individuals and those with T1DM is of utmost importance [10]. This insight may justify the incorporation of these variables into the models, potentially increasing their predictive accuracy.

Advances in sensor technologies have greatly simplified the real-time measurement of a myriad of physiological variables, which can be employed to estimate individual characteristics such as circadian phase [11] or directly utilized as inputs for data-driven models [12]. However, despite the existence of several open-source glucose prediction datasets and simulators, most of these resources predominantly focus on glucose levels and insulin administration, disregarding other physiological variables. This omission poses a significant challenge when attempting to model and gain a deeper understanding of the role of these variables in the glucose regulatory system. Notable examples of open-source datasets incorporating multiple physiological variables include the OhioT1DM dataset [13], and the more recent DiaTrend dataset [14]. Nevertheless, there is still a shortage of studies utilizing variables beyond glucose and insulin levels and attempting to determine their predictive capabilities, if any, for glucose prediction tasks in both healthy individuals and those with diabetes.

A rigorous data acquisition protocol and modeling approach are essential to determine the optimal combination of physiological variables, model architectures, and training strategies for blood glucose predictive modeling. To address these challenges, our study involves the collection of a novel dataset and systematic comparisons of several recurrent deep learning models, recognized for their robust performance in multi-step ahead glucose prediction, utilizing various combinations of input variables and diverse training methods. Our primary objective is to identify the model best suited for accurate glucose prediction in both healthy individuals and those with diabetes.

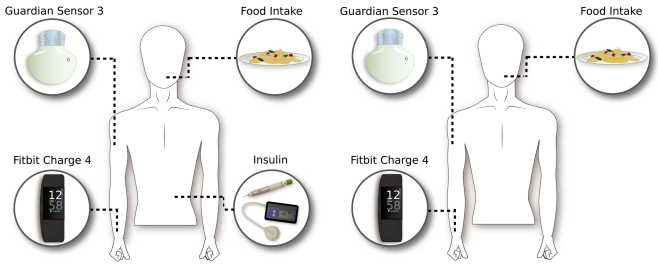

FIGURE 1. Monitoring devices used in the study.

The main contributions of this work are twofold: (i) we provide the research community with a novel dataset comprising a diverse set of physiological measurements collected from individuals with diabetes and those without the condition, all in real-world, free-living conditions, to promote and facilitate research in blood glucose prediction; and (ii) using this dataset, we conduct a comprehensive evaluation of several recurrent neural network architectures, trained with different approaches and using different sets of input variables, in order to address common questions that practitioners and researchers entering the field may have.

The remainder of the paper is organized as follows: Section II details the data collection process, data preprocessing steps, and the training methodologies employed for the selected models. Section III presents the evaluation results under various training techniques. Subsequently, Section IV offers a brief discussion and outlines the broader implications of our work. Finally, Section V provides concluding remarks.

II. MATERIALS AND METHODS

A. DATA COLLECTION

We collected ambulatory data from a total of 20 participants, consisting of both healthy individuals and those with T1DM. Recruitment was conducted through social media channels and flyers posted at the University. All participants provided informed written consent for their involvement in the study. The Scientific Ethics Committee of the School of Medicine, Pontificia Universidad Católica de Chile, granted approval for this study on May 5th, 2020, under Project 191015032. Each participant's involvement spanned seven days, during which they carried various ambulatory monitoring devices (Fig. 1). Specifically, we employed the Medtronic's Guardian Sensor 3 as a continuous glucose monitor and the Fitbit Charge 4 health and fitness tracker to capture heart rate and activity data. Additionally, participants maintained records of their dietary

TABLE 1. Monitored Physiological Variables and Sampling Periods

Time Series	Unit	Device	Sampling period
Glucose	mg/dl	Guardian sensor 3	5 minutes
Insulin	U	Report/Insulin pump	Variable
Heart rate	bpm	Fitbit Charge 4	5 seconds
Steps	-	Fitbit Charge 4	1 minute
Carbohydrates	grams	Report	Variable

consumption and insulin bolus administration through diary entries. For individuals with T1DM using insulin pumps, data on automatic insulin infusion were directly extracted from the pump reports. Importantly, the study was conducted within the framework of participants' normal daily activities, with no controlled schemes associated with carbohydrate intake or physical activity.

At the beginning of the study, we documented each volunteer's age, height, weight and self-reported sex. A health professional affixed the glucose sensor to either the back of the non-dominant upper arm or abdomen area of the volunteer. We verified the Bluetooth connection between the cellphone and the device and performed the first blood glucose calibration. Similarly, we conducted connection tests between the Fitbit band and its corresponding application. To prevent disruptions, all devices were fully charged before installation. Upon the study's conclusion, we extracted the data of interest and the number of carbohydrates and calories ingested were determined by an expert based on the participants' diet diaries.

B. DATA CONDITIONING

Before feeding the measured signals into the models for training, we standardized the sampling frequency for each signal to a uniform 5-minute interval, aligning with the highest sampling period among the variables detailed in Table 1. For variables originally recorded at a higher rate, we implemented a moving average technique to subsample the respective signal.

In addition, we discovered that better results could be achieved by smoothing impulsive inputs, namely, ingested carbohydrates and administered insulin. To this end, we applied physiological filters based on the Hovorka model [15]. This approach ensured that models received smooth and meaningful signals rather than the raw, impulsive data.

For carbohydrates, we employed the two-compartment gut absorption model [16], which characterized the digestion and absorption of carbohydrates with the following equations:

$$\frac{dD_1}{dt} = A_g c(t) - \frac{D_1(t)}{\tau_G} \tag{1}$$

$$\frac{dD_2}{dt} = \frac{D_1(t) - D_2(t)}{\tau_G} \tag{2}$$

$$IGAR(t) = \frac{D_2}{\tau_G} \tag{3}$$

Here, D_1 $[mmol]$ and D_2 $[mmol]$ represent the amount of glucose in compartments 1 and 2, respectively. A_g is an utilization factor of the absorption of carbohydrate to glucose,

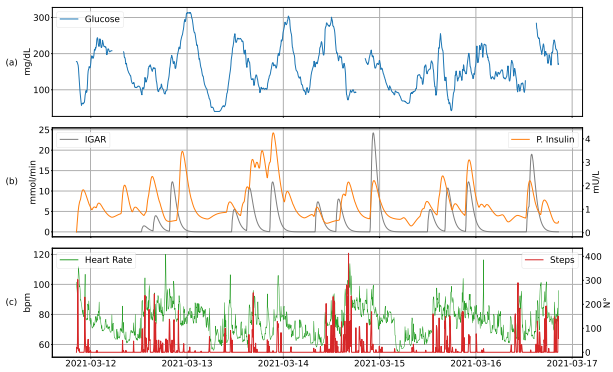

FIGURE 2. Processed signals of a representative subject with T1DM: (a) CGM; (b) IGAR and Plasma Insulin, and (c) Heart Rate and Number of Steps.

$c(t)$ $[mmol/min]$ is the amount of oral carbohydrate intake at any time expressed as glucose equivalents, and τ_G $[min]$ is the time of maximum appearance rate. $IGAR(t)$ [mmol/min] is the intestinal glucose absorption rate and is the signal fed into the models. In our study, we set $A_g = 0.8$, $\tau_G = 40[min]$, and assume that all carbohydrates from a given meal were uniformly ingested over a five-minutes period.

For insulin, we applied the Hovorka two-compartment model for insulin absorption, which is defined by the following equations:

$$\frac{dS_1}{dt} = u(t) - \frac{S_1(t)}{\tau_I} \tag{4}$$

$$\frac{S_2}{dt} = \frac{S_1(t) - S_2(t)}{\tau_I} \tag{5}$$

$$\frac{dI}{dt} = \frac{S_2(t)V_I}{\tau_I} - k_e I(t) \tag{6}$$

In these equations, $u(t)$ represents the administered insulin rate (combining bolus and basal) in $[mU/min]$, τ_I is the time of maximum appearance rate in $[min]$, V_I denotes the insulin distribution volume in $[L]$, k_e stands for the insulin elimination rate in $[1/min]$, and I corresponds to the plasma insulin concentration in $[mU/L]$. The plasma insulin concentration signal is fed into the models. For our study, we utilized $\tau_I = 55[min]$, $k_e = 0.138[1/min]$ and $V_I = 0.12 * BW[L]$, where BW is the individual's body weight in kilograms.

Finally, in cases of missing values due to glucose sensor failure or the volunteer removing the health tracker, we limited our interpolations to gaps no longer than 12 samples (equivalent to one hour when $T = 5$ minutes). We observed that longer interpolations tended to produce unrealistic results. Fig. 2 displays all the processed signals for a representative individual with diabetes.

C. MODELING DETAILS

For modeling purposes, we conducted a comparative analysis of several recurrent neural network architectures. The implemented models vary in complexity, input signal combinations, and training methodology (population, personalized training,

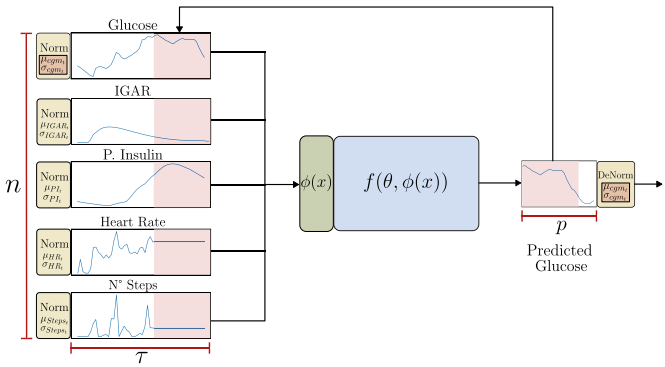

FIGURE 3. Inference process common to all models. The architecture depicted includes all the available variables, which can vary based on the selected group (see Table 2). Within each n-dimensional sliding window of length τ, input variables are normalized across the time dimension. A preprocessing step involving the feature extraction layer ϕ is applied to the inputs, which are then processed by the primary network to produce a one-step-ahead prediction. This prediction is added to the moving window as if it were a new measurement. Exogenous variables are propagated as described in Section II-C. This iterative process is repeated until the requested p predictions are generated. The predicted sequence is then denormalized using parameters μ_{cgm_t} and σ_{cgm_t}, obtained during the normalization.

and fine-tuning). This exploration aims to identify the most suitable architecture, input variable combination, and training approach for multi-step ahead glucose prediction in both healthy individuals and those with diabetes.

For each model, with parameters denoted as θ, we employ a recursive prediction strategy, as shown in Fig. 3. An n-dimensional sliding window of length τ, comprising past measurements, serves as input to generate a one-step-ahead prediction. This prediction is then integrated into the sliding window as if it were a new glucose measurement, facilitating the generation of the next prediction. This iterative process continues until the desired p predictions are obtained.

The handling of exogenous variables (when used) varies depending on their nature. Specifically, the last measured values of heart rate and the number of steps are propagated as estimates of their future values. Conversely, for IGAR and plasma insulin, we apply Hovorka's model equations, as described earlier, to extrapolate their values over the prediction horizon.

Additionally, for each model, we incorporate a feature extraction layer ϕ preceding any recurrent layers. We also conduct a min-max normalization across the time dimension of all the input sequences, applying a linear transformation to the original data and mapping it to the range (0, 1) before feeding it into the network. This practice has been observed to enhance overall performance in this task.

The selected evaluation metric for all the models is the root mean squared error (RMSE), which is assessed at the last point of the predicted sequence. Specifically, it measures the accuracy of the predicted glucose value p minutes ahead:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Glucose_{CGM,i} - Glucose_{pred,i}\right)^2}, \quad (7)$$

where $Glucose_{CGM,i}$ denotes the experimental glucose level measured by the CGM system at time i, and $Glucose_{pred,i}$ represents the glucose predicted by the model for time i. It is important to note that predictions are projected 30 minutes into the future, utilizing only information from the past. In our context, RMSE is expressed in glucose units (mg/l), with the optimal value being zero, signifying a perfect prediction. This metric selection enables relevant comparisons with other studies, given that RMSE is a widely reported metric in this field.

In all experiments, both training and evaluation were conducted with a prediction horizon of $p = 30$ minutes, using information from the past $\tau = 2$ hours. This setup aligns with the standards commonly used in the field and ensures consistency for comparative purposes.

To obtain robust and reliable results, a cross-validation procedure with 5-folds was employed. This approach helps mitigate the impact of data variability. The data was split into training, validation, and testing sets with a ratio of 70-15-15, respectively, for each individual participant.

The results reported in this study are presented as the average values, along with their corresponding standard deviations, obtained from the five folds of the cross-validation process. This practice ensures that the reported outcomes accurately represent the performance of the predictive models under varying conditions.

Statistical comparisons were made using Friedman's test and Tukey's test for post-hoc analysis to evaluate the performance differences among the various recurrent neural network architectures, groups of variables, and training methodologies, considering a significance level of 0.05.

D. NEURAL ARCHITECTURES

The neural architectures were chosen due to their proven effectiveness in modeling time series data and their state-of-the-art performance in glucose prediction [17], [18]. In the following, brief descriptions of all the implemented neural architectures are provided:

1) LONG SHORT-TERM MEMORY (LSTM)

The first and simplest implemented architecture comprises a Long Short-Term Memory (LSTM) layer following the feature extraction layer, with two additional feed-forward layers at the end for converting LSTM features into the output space.

2) ENCODER-DECODER (ENC-DEC)

In this configuration, the network consists of an encoder that takes input features from the feature extraction layer and transforms them into a latent intermediate vector using an LSTM layer. Subsequently, the decoder takes these latent vectors and transforms them into the output space using a second LSTM layer and two feed-forward layers on top of it.

TABLE 2. Groups of Variables Used for Training the Neural Networks

Variable	G1	G2	G3	G4	G5
CGM	X	X	X	X	X
IGAR		X	X	X	
P. Insulin		X	X	X	
H.R.			X	X	X
Steps				X	X

Insulin is Only Considered for TDM Subjects.

3) BIDIRECTIONAL ENCODER-DECODER (BI ENC-DEC)

This network shares the same architecture as the Encoder-Decoder configuration but includes a bidirectional LSTM layer [19] at the encoder, which processes the input sequences in both forward and backward directions.

4) ENCODER-DECODER WITH DOUBLE ATTENTION (ENC-DEC DATTN)

This architecture, initially introduced in [20] for multivariable time series prediction tasks, features an Encoder-Decoder design equipped with variable attention in the encoder and temporal attention in the decoder. The variable attention mechanism assigns weights to each input variable or feature based on its contribution to the prediction. Subsequently, the temporal attention dynamically weights the produced hidden states of the encoder LSTM layer, enabling the network to focus on different temporal segments of the input sequences that are relevant for the prediction. This design provides greater flexibility to the network, making it more effective, especially for longer input sequences.

E. GROUPS OF VARIABLES

For each of the model architectures described above, we considered five groups of variables to train multiple glucose prediction models. This approach enables us to compare the performance of these models when different sets of variables are available. For individuals with T1DM, insulin was considered alongside IGAR because both data sources are derived from self-reports or the insulin pump when available. Therefore, it appears reasonable that if a subject can report food consumption, they should also be able to register insulin infusion. The composition of each variable group is detailed in Table 2.

F. MODELING APPROACHES

Additionally, all model architectures, using the various combinations of input variables, were trained using different approaches: population training, personalized training, and fine-tuning training. A detailed overview of these methodologies is presented below:

1) POPULATION TRAINING

We initiated our evaluation by assessing the performance of population-based models in predicting future glucose values for individuals within their respective populations. These models were trained using the complete training data for either the healthy or T1DM populations. Subsequently, they were tested on individual test sets of each participant. This experiments were conducted with a learning rate of 1×10^{-3} and a prediction horizon of 30 minutes. In a clinical context, this approach would imply employing the same set of model weights to predict future glucose values for all patients. In the event of incorporating a new individual, their data would be integrated into the training dataset, prompting the retraining and updating of the population model, which would then be applied to all individuals.

2) PERSONALIZED TRAINING

In contrast, personalized training encompasses the training a different model for each individual within the population, followed by an evaluation on the respective individual's test set. This approach has the evident advantage of insulating model performance from the inter-subject variability of the population. However, as we will delve into later, the performance of this approach is highly dependent on the amount and quality of the data available for a specific individual.

3) FINE-TUNING TRAINING

Finally, we assessed the performance of models fine-tuned to a particular individual. In this approach, a population-based model trained with all the data of $n - 1$ individuals (where n signifies the population size) is subject to fine-tuning using the training data of the target individual. The model's performance is subsequently evaluated using the individual's testing data. During the fine-tuning (as well as the personalized) training process, the learning rate is reduced to 1×10^{-4}.

G. PERFORMANCE ASSESSMENT

In addition to RMSE, we employ the Clarke Error Grid to evaluate the clinical significance of discrepancies between predicted and actual glucose values. This grid is segmented into five distinct zones, each indicating varying levels of clinical relevance. Zone A represents clinically accurate predictions, where the predicted values falling within $\pm 20\%$ of the actual values would lead to similar clinical outcomes if used for decision-making. Predictions in Zone B, though less accurate, remain within a safe margin. Falling outside the $\pm 20\%$ range, they do not prompt opposite clinical decisions but may warrant further investigation or action. Zone C signifies predictions with discrepancies that could potentially lead to over- or under-treatment if solely relied upon for clinical decisions. Zone D indicates predictions with dangerous discrepancies. Despite falling within the normal range, reliance on these values could result in significant errors in clinical actions. Predictions falling into Zone E display values entirely opposite to the actual ones, suggesting errors that would likely prompt opposite clinical actions.

The use of the Clarke Error Grid enables the visual assessment of a glucose monitoring system or prediction model. This tool evaluates not only the predictions' accuracy but also discerns the potential clinical implications of any inaccuracies. It aids in identifying the predominant zones where data

TABLE 3. Characteristics of the Dataset Obtained From the Clinical Trial

Characteristic	Value
Participants (n)	20
Age (mean/SD)	27.2 (4.0)
BMI (mean/SD)	26.1 (5.9)
Women (%)	50.0
T1DM (%)	45.0 (9 individuals)
Steps/day (mean/SD)	8,579 (4,050)
CV_{HT} (mean/SD)	0.15 (0.02)
CV_{T1DM} (mean/SD)	0.34 (0.09)
[5th-95th] CGM $Percentiles_{HT}$ (mg/dl)	[70.9, 117.1]
[5th-95th] CGM $Percentiles_{T1DM}$ (mg/dl)	[66.0, 248.4]

BMI stands for body mass index, and CV represents the coefficient of variation [22], a measurement of glycemic variability. All averages and standard deviations are computed across the participant population.

points fall, providing insights into the overall accuracy and clinical relevance of the system.

III. RESULTS

Study participants were divided into two groups: healthy individuals (estimated glycosylated hemoglobin (HbA1c) [21] between 4.39% and 4.94%) and those with T1DM (estimated HbA1c between 6.05% and 8.21%). The inclusion of healthy subjects in this study aimed to identify the most relevant variables for predicting glycemia, not only for individuals with diabetes but also for those at risk of developing the disease. Table 3 provides key characteristics of the participants.

An interesting initial finding in our dataset is that the energy intake obtained from the food diaries was approximately 15% lower than the resting energy expenditure estimated from the Mifflin-St Jeor Equation [23]. This observation may be attributed to the systematic under-reporting of energy intake, a phenomenon often observed with self-reported dietary instruments under free-living conditions [24].

The dataset, named UCHTT1DM, includes data collected from 20 participants, comprising information on glucose, heart rate, IGAR, steps, consumed carbohydrates, and insulin (provided only for subjects with T1DM). This dataset is made available to the community,[1] offering comprehensive physiological information from both healthy individuals and those with T1DM.

A. EFFECT OF MODEL ARCHITECTURES, INPUT VARIABLES AND MODELING APPROACHES

The results are summarized in Fig. 4 for both healthy individuals and patients with T1DM. From this figure, it becomes evident that the predictive performance of models for healthy individuals is considerably better than for individuals with diabetes. This discrepancy is likely due to the lower glycemic variability and fewer extreme values observed in the healthy population (as indicated in Table 3), which are easier to predict with data-driven models.

[1][Online]. Available: https://github.com/fisiologiacuantitativauc/UC_HT_T1DM

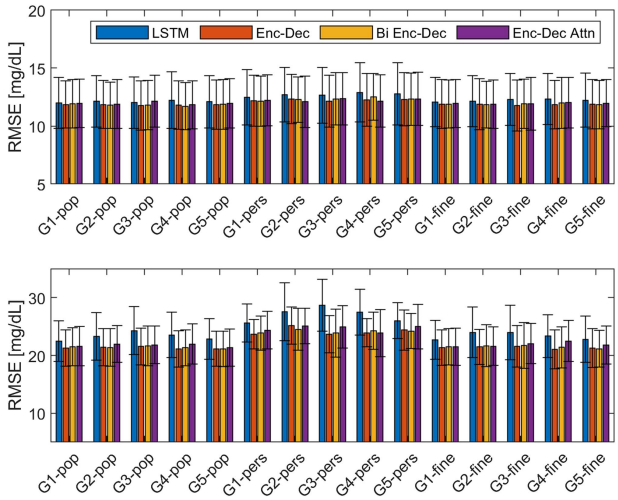

FIGURE 4. Results for the healthy and T1DM population. Here *pop*, *pers*, and *fine*, stand for the population, personalized, and fine-tuning training approaches, respectively.

TABLE 4. RMSE [mg/dL] At 30 Minutes Ahead Predictions for the Population Training Approach

Healthy Subjects				
Group\Model	LSTM	Enc-Dec	Bi Enc-Dec	Enc-Dec DAttn
G1	12.010 (0.92)	**11.861 (0.86)**	11.917 (0.93)	11.89 (0.85)
G2	12.131 (1.00)	11.858 (0.92)	**11.795 (0.88)**	11.949 (0.91)
G3	12.029 (0.96)	**11.773 (0.94)**	11.801 (0.97)	11.803 (0.88)
G4	12.237 (1.07)	11.809 (0.98)	**11.714 (0.92)****	11.885 (0.90)
G5	12.097 (1.05)	**11.839 (0.89)**	11.868 (0.95)	11.943 (0.91)
Subjects with T1DM				
Group\Model	LSTM	Enc-Dec	Bi Enc-Dec	Enc-Dec Attn
G1	23.460 (2.08)	22.130 (1.99)	22.255 (1.87)	**22.128 (1.82)**
G2	24.550 (2.62)	**22.151 (2.15)**	22.240 (2.09)	22.219 (1.92)
G3	25.602 (3.22)	22.345 (2.15)	22.702 (2.47)	**22.188 (1.89)**
G4	25.398 (3.31)	**22.008 (1.99)****	22.378 (2.12)	22.324 (1.77)
G5	24.078 (2.43)	22.055 (1.85)	**22.032 (1.91)**	22.154 (1.84)

The best results across the different model architectures for the same group of input variables are shown in bold. The best result in the table is indicated with two asterisks (**).

1) POPULATION TRAINING

The results of population training, are summarized in Table 4. Despite minimal differences in prediction among the different groups of input variables, we can observe slightly superior results for Group 4 (G4) in healthy individuals. The best-performing model within this group is the Bidirectional Encoder-Decoder architecture. For individuals with diabetes, the differences between input variable groups are even less pronounced, suggesting that the contribution of exogenous variables, at least with the selected model architectures, has an almost negligible impact on predicting future glucose values for this population.

When comparing various model architectures, all Encoder-Decoder architectures exhibit similar performance, significantly outperforming the LSTM architecture, which consistently delivers poorer results across all the input variables groups, both for the healthy and diabetic populations. Notably, the Encoder-Decoder with Double Attention displays less variation in performance across individuals, resulting in a smaller standard deviation. This consistency may be attributed

TABLE 5. RMSE [mg/dL] At 30 Minutes Ahead Predictions for the Personalized Training Approach

Healthy Subjects				
Group\Model	LSTM	Enc-Dec	Bi Enc-Dec	Enc-Dec Attn
G1	12.484 (1.14)	12.183 (0.92)	**12.150 (0.91)**	12.189 (0.98)
G2	12.719 (1.29)	12.339 (1.02)	12.284 (0.80)	**12.215 (1.01)**
G3	12.651 (1.12)	12.156 (0.98)	12.346 (1.01)	**12.131 (0.97)****
G4	12.905 (1.40)	12.255 (0.96)	12.511 (0.92)	**12.227 (1.02)**
G5	12.776 (1.28)	12.309 (0.93)	12.316 (0.81)	**12.193 (0.98)**
Subjects with T1DM				
Group\Model	LSTM	Enc-Dec	Bi Enc-Dec	Enc-Dec Attn
G1	30.159 (5.35)	**28.185 (6.02)**	28.588 (5.65)	28.812 (6.30)
G2	31.715 (5.64)	29.695 (5.24)	29.043 (6.44)	**27.898 (6.84)****
G3	32.946 (5.89)	**28.429 (6.19)**	28.466 (5.71)	28.583 (6.16)
G4	31.664 (6.22)	28.650 (5.71)	28.903 (5.76)	**28.334 (5.87)**
G5	30.519 (6.49)	**28.948 (5.16)**	29.020 (6.39)	29.388 (5.14)

The best results across the different model architectures for the same group of input variables are shown in bold. The best result in the table is indicated with two asterisks (**).

TABLE 6. RMSE [mg/dL] At 30 Minutes Ahead Predictions for the Fine-Tuning Training Approach

Healthy Subjects				
Group\Model	LSTM	Enc-Dec	Bi Enc-Dec	Enc-Dec Attn
G1	12.06 (1.01)	11.903 (0.93)	**11.901 (0.93)**	11.945 (0.95)
G2	12.139 (0.93)	11.881 (0.97)	**11.845 (0.91)**	11.888 (0.97)
G3	12.282 (1.08)	**11.780 (0.94)****	11.926 (1.02)	11.919 (0.99)
G4	12.341 (0.98)	**11.854 (0.93)**	11.998 (0.94)	12.018 (1.02)
G5	12.231 (1.01)	11.888 (0.97)	**11.836 (0.81)**	11.972 (0.96)
Subjects with T1DM				
Group\Model	LSTM	Enc-Dec	Bi Enc-Dec	Enc-Dec Attn
G1	23.713 (2.22)	22.036 (1.88)	22.15 (1.82)	**21.987 (1.84)**
G2	25.016 (2.41)	**22.357 (1.93)**	22.428 (1.85)	22.650 (2.03)
G3	25.011 (2.57)	**22.461 (2.25)**	22.766 (2.13)	22.995 (2.02)
G4	24.545 (2.60)	**22.559 (2.40)**	22.827 (2.17)	23.490 (2.37)
G5	23.885 (2.52)	22.070 (2.13)	**21.929 (1.88)****	22.408 (2.06)

The best results across the different model architectures for the same group of input variables are shown in bold. The best result in the table is indicated with two asterisks (**).

to the effect of the attention mechanism, which appears to have learned to compensate for the inter-individual differences, a phenomenon that warrants further investigation.

2) PERSONALIZED TRAINING

Here, we present the results of personalized training. As depicted in Table 5, the average performance among the healthy individuals is only slightly worse than for the other approaches, owing to the substantial data available for each subject. Conversely, for individuals with T1DM, the average performance significantly diminishes due to certain patients having limited data, notably impacting the overall results (e.g., T1DM subject N° 9).

Across various groups, the performance differences are minimal. Yet, it is notable that the Encoder-Decoder with Double Attention consistently stands out for both healthy and diabetic populations. This prominence could be attributed to the attention mechanism, which, despite not being fully optimized due to data limitations, aids in compensating for variations and unmeasured disruptions within the targeted individuals.

3) FINE-TUNING TRAINING

Table 6 shows the results of the fine-tuning evaluation for the healthy and diabetic populations. As before, the difference in performance between the groups of input variables is almost negligible. However, it can be seen that both the Encoder-Decoder and the Bidirectional Encoder-Decoder show greater performance advantages over the Encoder-Decoder with Double Attention when compared to the population training results. This is probably due to the fact that the latter needs more data of a particular individual to better re-adjust the attention weights.

B. COMPARISON OF DIFFERENT TRAINING APPROACHES

Fig. 4 illustrates the variation in RMSE at 30 minutes ahead predictions across different input variables groups and training approaches. While the mean values do not exhibit significant variation across the groups for the same training approach, notable differences emerge when comparing the different training approaches.

Firstly, it is evident that, in the case of the T1DM population, models using the personalized training approach show a large variability, as indicated by the long error bars. This variability is due to the fact that personalized models are highly sensitive to the amount of data available for each individual. Subjects with a substantial amount of training data perform much better than individuals with a modest amount of training data, leading to significant performance differences. As seen in Fig. 4, this is not the case for the population or fine-tuning training approaches. This underscores the critical importance of avoiding personalized models when dealing with limited individual data, favoring population or fine-tuned models in such scenarios.

When comparing population and fine-tuning approaches, although population models tend to perform slightly better on average, fine-tuned models exhibit less variability in prediction errors, as indicated by the shorter error bars. This characteristic is of paramount importance in safety-critical applications like glucose prediction and control, where significant errors can have severe, even life-threatening consequences. Therefore, models with reduced variability in prediction errors should be prioritized.

Remarkably, among all groups, models in G1 utilizing the fine-tuning training approach demonstrate the least variability. Notably, the database comprises individuals with an average daily step count of 8,579, falling within the category of "somewhat active" as per the classifications established by Tudor-Locke and Bassett [25]. This places the cohort within the moderately active lifestyle category defined by the graduated step index. This observation suggests that incorporating additional variables, such as step count, might yield more pronounced

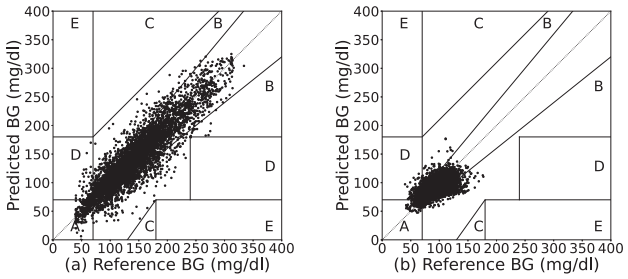

FIGURE 5. Clarke Error Grids representing the clinical significance of predicted versus actual blood glucose (BG) values, categorized into five distinct zones A to E, aiding in the visual assessment of glucose prediction accuracy and potential clinical implications. (a) patients with T1DM, (b) healthy subjects.

benefits for individuals with higher activity levels compared to those who are less active.

C. CLINICAL RELEVANCE OF THE PREDICTIONS

The Clarke Error Grids depicted in Fig. 5 highlight compelling insights derived from the best performing model. The assessment indicates that the utilization of this model in treatment plans significantly mitigates the risk of encountering hyperglycemia-associated zones. In the case of patients with T1DM, the model maintains a reassuring distribution with 80.4% of readings falling within zone A, denoting clinically accurate predictions, while 18.1% are within zone B, indicating predictions that, though less accurate, still suggest safe outcomes. Importantly, less than 2% of the predictions fall outside these clinically acceptable ranges, demonstrating a high level of safety and accuracy. Conversely, for healthy individuals, the model performs even more robustly, with 88.5% of predictions in zone A, showcasing the accuracy in these readings, and 9.4% falling within zone B, which still implies safe outcomes. A minimal 2% of predictions fall within zone D, representing discrepancies that, while indicating danger, remain within the normal range, underlining the model's ability to maintain a high level of accuracy and safety, particularly in predicting glucose levels for healthy individuals.

IV. DISCUSSION

The results and comparisons across different modeling axes (individual condition, model architecture, input variables, and training size) demonstrate the effectiveness of deep learning methods in building population-based and personalized models for glucose prediction using data obtained under free-living conditions. To the best of our knowledge, this is the first work to address all these axes and provide insights for practitioners intending to use deep learning models for glucose prediction.

When comparing the results between the populations of individuals with diabetes and those without the condition, it was found that the models for the healthy population significantly outperformed all the models for the diabetic population, regardless of the model architecture and groups of input variables available. This discrepancy is likely due to the lower inter-subject variability in the healthy population, lower glycemic variability for each individual, and the tendency of healthy subjects to exhibit fewer extreme values (see Table 3), which are challenging to predict accurately with data-driven models.

Regarding different model architectures, we observed that all the Encoder-Decoder models consistently outperformed the LSTM architecture. Among the Encoder-Decoder architectures, the simple Encoder-Decoder showed the best results overall, possibly due to its simplicity and the limited amount of data available to better train more complex models. However, the Encoder-Decoder with Double Attention exhibited interesting features, including less inter-subject variability in population training and the best performance in personalized training. We suspect that better performance with this architecture can be obtained when more training data is available.

Contrary to expectations, no significant improvements were found when incorporating exogenous inputs for predicting future glucose values with the tested architectures. Although Group 4 (which contained all input variables) seemed to have slightly better performance than other groups, it exhibited a wider RMSE distribution than Group 1, making it less suitable for this safety-critical application. For future work, other architectures more heavily focused on attention, such as transformer-based architectures [26], [27], which have shown promising results in multivariate time series forecasting, will also be evaluated to determine if the information carried in the exogenous variables can be utilized more effectively.

Finally, when comparing different training approaches, we found that the personalized approach generally produced the least favorable results. This is primarily due to its sensitivity to variations in the amount of data available for each individual, resulting in wider error distributions and greater performance variation across individuals. In contrast, the fine-tuning approach generally yielded narrower error distributions and less variation in performance across individuals, even when the population training approach achieved better average performances. Notably, the fine-tuning approach demonstrated remarkable robustness to variations in the amount of data for each individual and effectively transferred knowledge to new subjects, even with minimal available data. For future work, exploring other training methods such as meta-learning, which has shown promising performance in glucose prediction [17], will be a valuable avenue of investigation.

V. CONCLUSION

In this work, we leveraged deep learning methods to develop personalized data-driven models for blood glucose prediction, addressing both patients with T1DM and healthy subjects, while comparing them across several critical dimensions: model architectures, input variables, training approaches, and variations in training data size. Our goal was to provide valuable insights to practitioners and researchers seeking to apply deep learning techniques for glucose prediction in diverse

applications, including but not limited to alarm systems, artificial pancreas development, and automatic bolus calculators.

Among the various training approaches examined, the fine-tuning approach emerged as the most effective and robust method, particularly in scenarios where data availability varies among individuals. Additionally, across the different groups of input variables tested, models trained exclusively on past glucose values exhibited a slight performance edge and produced narrower error distributions, a crucial factor in safety-critical applications like blood glucose prediction. Furthermore, the simple Encoder-Decoder architecture demonstrated its effectiveness, outperforming more complex models given the limited available data. However, the Encoder-Decoder with Double Attention displayed competitive results, suggesting potential improvements in datasets with larger data volumes.

Future investigation in this domain will encompass the exploration of alternative model architectures, such as transformers, which heavily rely on attention mechanisms and may better utilize exogenous variables [26], [27]. Additionally, exploring and comparing alternative training approaches, including meta-learning, is a promising avenue given its demonstrated potential in recent related research [17]. Furthermore, comparative analysis involving other well-known datasets in the field, such as the OhioT1DM dataset [13], will provide valuable insights into the generalizability of our findings.

To foster further research in this area, we have anonymized and made our curated dataset publicly available.[2] We encourage fellow researchers to access and leverage this resource for their own investigations and advancements in glucose prediction.

REFERENCES

[1] G. Roglic et al., "WHO global report on diabetes: A summary," *Int. J. Noncommunicable Dis.*, vol. 1, no. 1, pp. 3–8, 2016.

[2] P. Saeedi et al., "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the international diabetes federation diabetes atlas," *Diabetes Res. Clin. Pract.*, vol. 157, 2019, Art. no. 107843.

[3] T. Battelino et al., "Clinical targets for continuous glucose monitoring data interpretation: Recommendations from the international consensus on time in range," *Diabetes Care*, vol. 42, no. 8, pp. 1593–1603, 2019.

[4] C. Gorst et al., "Long-term glycemic variability and risk of adverse outcomes: A systematic review and meta-analysis," *Diabetes Care*, vol. 38, no. 12, pp. 2354–2369, 2015.

[5] L. M. Huyett, E. Dassau, H. C. Zisser, and F. J. Doyle, "Glucose sensor dynamics and the artificial pancreas: The impact of lag on sensor measurement and controller performance," *IEEE Control Syst. Mag.*, vol. 38, no. 1, pp. 30–46, Feb. 2018.

[6] S. Mehmood, I. Ahmad, H. Arif, U. E. Ammara, and A. Majeed, "Artificial pancreas control strategies used for type 1 diabetes control and treatment: A comprehensive analysis," *Appl. Syst. Innov.*, vol. 3, no. 3, 2020, Art. no. 31.

[7] M. M. Seron, A. M. Medioli, and G. C. Goodwin, "A methodology for the comparison of traditional MPC and stochastic MPC in the context of the regulation of blood glucose levels in type 1 diabetics," in *Proc. Australian Control Conf.*, 2016, pp. 126–131.

[8] E. Montaser, J.-L. Díez, and J. Bondia, "Glucose prediction under variable-length time-stamped daily events: A seasonal stochastic local modeling framework," *Sensors*, vol. 21, no. 9, 2021, Art. no. 3188.

[9] T. Diamond, F. Cameron, and B. W. Bequette, "A new meal absorption model for artificial pancreas systems," *J. Diabetes Sci. Technol.*, vol. 16, no. 1, pp. 40–51, 2022.

[10] S. Bahremand, H. S. Ko, R. Balouchzadeh, H. Felix Lee, S. Park, and G. Kwon, "Neural network-based model predictive control for type 1 diabetic rats on artificial pancreas system," *Med. Biol. Eng. Comput.*, vol. 57, no. 1, pp. 177–191, 2019.

[11] A. Suarez, F. Nunez, and M. Rodriguez-Fernandez, "Circadian phase prediction from non-intrusive and ambulatory physiological data," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 5, pp. 1561–1571, May 2021.

[12] T. Zhu, C. Uduku, K. Li, P. Herrero, N. Oliver, and P. Georgiou, "Enhancing self-management in type 1 diabetes with wearables and deep learning," *NPJ Digit. Med.*, vol. 5, no. 1, pp. 1–11, 2022.

[13] C. Marling and R. Bunescu, "The OhioT1DM dataset for blood glucose level prediction: Update2020," *CEUR Workshop Proc.*, vol. 2675, pp. 71–74, 2020.

[14] T. Prioleau, A. Bartolome, R. Comi, and C. Stanger, "DiaTrend: A dataset from advanced diabetes technology to enable development of novel analytic solutions," *Sci. Data*, vol. 10, no. 1, Aug. 2023, Art. no. 556.

[15] R. Hovorka et al., "Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes," *Physiol. Meas.*, vol. 25, no. 4, 2004, Art. no. 905.

[16] D. Boiroux, D. A. Finan, J. B. Jørgensen, N. K. Poulsen, and H. Madsen, "Optimal insulin administration for people with type 1 diabetes," *IFAC Proc. Volumes*, vol. 43, no. 5, pp. 248–253, 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1474667016300428

[17] S. Langarica, M. Rodriguez-Fernandez, F. Núñez, and F. J. Doyle, "A meta-learning approach to personalized blood glucose prediction in type 1 diabetes," *Control Eng. Pract.*, vol. 135, 2023, Art. no. 105498.

[18] M. F. Rabby, Y. Tu, M. I. Hossen, I. Lee, A. S. Maida, and X. Hei, "Stacked LSTM based deep recurrent neural network with Kalman smoothing for blood glucose prediction," *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 1, Mar. 2021, Art. no. 101.

[19] Z. Cui, R. Ke, and Y. Wang, "Deep stacked bidirectional and unidirectional LSTM recurrent neural network for network-wide traffic speed prediction," in *Proc. ACM SIGKDD Int. Workshop Urban Comput. (UrbComp)*, 2017.

[20] Y. Qin, D. Song, H. Chen, W. Cheng, G. Jiang, and G. W. Cottrell, "A dual-stage attention-based recurrent neural network for time series prediction," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 2627–2633.

[21] D. M. Nathan, J. Kuenen, R. Borg, H. Zheng, D. Schoenfeld, and R. J. Heine, "Translating the A1C assay into estimated average glucose values," *Diabetes Care*, vol. 31, no. 8, pp. 1473–1478, 2008.

[22] D. Czerwoniuk, W. Fendler, L. Walenciak, and W. Mlynarski, "GlyCulator: A glycemic variability calculation tool for continuous glucose monitoring data," *J. Diabetes Sci. Technol.*, vol. 5, no. 2, pp. 447–451, Mar. 2011.

[23] M. D. Mifflin, S. T. St Jeor, L. A. Hill, B. J. Scott, S. A. Daugherty, and Y. O. Koh, "A new predictive equation for resting energy expenditure in healthy individuals," *Amer. J. Clin. Nutr.*, vol. 51, no. 2, pp. 241–247, 1990.

[24] M. N. Ravelli and D. A. Schoeller, "Traditional self-reported dietary instruments are prone to inaccuracies and new approaches are needed," *Front. Nutr.*, vol. 7, 2020, Art. no. 90.

[25] C. Tudor-Locke and D. R. Bassett, "How many steps/day are enough? Preliminary pedometer indices for public health," *Sports Med.*, vol. 34, pp. 1–8, 2004.

[26] Q. Wen et al., "Transformers in time series: A survey," in *Proc. 32nd Int. Joint Conf. Artif. Intell.*, 2023, pp. 6778–6786.

[27] B. Lim, S. O. Arik, N. Loeff, and T. Pfister, "Temporal Fusion Transformers for interpretable multi-horizon time series forecastingTemporal fusion transformers for interpretable multi-horizon time series forecasting," *Int. J. Forecasting*, vol. 37, no. 4, pp. 1748–1764, 2021.

[2][Online]. Available: https://github.com/fisiologiacuantitativauc/UC_HT_T1DMonline